Prasiddha Narayan Saurabh

#1. Write a query to calculate what % of the customers have made a claim in the current exposure period[i.e. in the given dataset]?

- Around 5% of the customers have made claim at least once. (total customers: 678013, positive claims: 34060, percentage of customers making a claim: 5.02%)

#2.

2.1. Create a new column as 'claim_flag' in the table 'auto_insurance_risk' as integer datatype.

2.2. Set the value to 1 when ClaimNb is greater than 0 and set the value to 0 otherwise.

- Done in Sql file

#3.

3.1. What is the average exposure period for those who have claimed?

3.2. What do you infer from the result?

- Average exposure period is higher for those who have made a claim at least once

#4.

4.1. If we create an exposure bucket where buckets are like below, what is the % of total claims by these buckets?

4.2. What do you infer from the summary?

- 45% of the total claims come from 4th bucket(highest exposure). That is higher exposure translates higher probability of an accident.

#5. Which area has the highest number of average claims? Show the data in percentage w.r.t. the number of policies in the corresponding Area.

- Area F has highest number of average claims (i.e., percentage of claims with respect to number of policies taken)

#6. If we use these exposure bucket along with Area i.e. group Area and Exposure Buckets together and look at the claim rate, an interesting pattern could be seen in the data. What is that?

- claim rate increases with increasing exposure in every area. Also, average claim rate increases with area from A towards F

#7.

7.1. If we look at average Vehicle Age for those who claimed vs those who didn't claim, what do you see in the summary?

- Average vehicle age is shorter for those who claimed compared to those who didn’t. This might be due to the fact that customers with older vehicles (higher vehicle age) are more versed with the controls of their vehicle while driving compared to those with newer vehicles.

7.2. Now if we calculate the average Vehicle Age for those who claimed and group them by Area, what do you see in the summary? Any particular pattern you see in the data?

- Average vehicle age (for those who claimed) decreases from A to F. which may either mean that people of area F frequently replace their vehicles with new ones (case for a developed economy) or they are recently growing economically so as to be able to buy vehicles for the first time(case for a developing economy)

#8. If we calculate the average vehicle age by exposure bucket(as mentioned above), we see an interesting trend between those who claimed vs those who didn't. What is that?

- In every exposure bucket, the average age of vehicles for those who claim is smaller than those who didn't. Also, average age of vehicles increases with exposure bucket (obviously)

#9.
9.1. Create a Claim_Ct flag on the ClaimNb field as below, and take the average of the BonusMalus by Claim_Ct.
9.2. What is the inference from the summary?

- average bonusmalus is highest for those who made multiple claims followed by single claim and subsequently no claim. This result is as expected i.e., customers making more claims are subjected to higher bonusmalus
- If we group by absolute number of claims, we see that customers making 4 claims or higher are rare. However, their bonus malus is infact lower than those who claimed thrice.

#10. Using the same Claim_Ct logic created above, if we aggregate the Density column (take average) by Claim_Ct, what inference can we make from the summary data?

- We can see that average density from areas where multiple claims have been filed is highest followed by single claim and no claim. The summary reveals that higher density translates to higher claims (i.e., higher accidents), thus the customers can be charged premiums for being in high density areas.

#11. Which Vehicle Brand & Vehicle Gas combination have the highest number of Average Claims (use ClaimNb field for aggregation)?

- vehicle B12 on regular gas has the highest number of claims as well as highest % of claims on total number policies. As defined in question 5: average claim=% of claims (in the segment) from total number of policies(in the segment)

#12. List the Top 5 Regions & Exposure[use the buckets created above] Combination from Claim Rate's perspective. Use claim_flag to calculate the claim rate.

- Region: Exposure bucket: Claim rate

R42	3	7.8
R82	4	7.6
R11	4	7.5
R53	4	7.4
R25	4	7.3

#13.
13.1. Are there any cases of illegal driving i.e. underaged folks driving and committing accidents?

- Nope there no cases of driver being below 18 years of age

13.2. Create a bucket on DrivAge and then take the average of BonusMalus by this Age Group Category. WHat do you infer from the summary?

- Average bonusmalus decreases with age. This means that older drivers are supposed to be more experienced and are subjected to lesser bonusmalus.
- However, on viewing avg claim rate (acr) w.r.t driver age, acr decreases during 18 years till late 20s and again starts increasing after 60 years of age. Drivers in senior category are claiming higher without being subjected to higher bonusmalus

Conceptual

#14. Mention one major difference between unique constraint and primary key?

- Primary key identifies each row of a table and thus it can never be null. Unique constraint can be null. We can have multiple unique constraints.

#15. If there are 5 records in table A and 10 records in table B and we cross-join these two tables, how many records will be there in the result set?

- 50 records

#16. What is the difference between inner join and left outer join?

- In inner join, all the entries pertaining to the common identifiers (identifier should be present in both the tables) are returned. In left outer join, returns all the entries pertaining to common identifier (i.e., identifier that is present in both the tables) and also returns all other entries from left table even if no common identifier has been found in right table.

#17. Consider a scenario where Table A has 5 records and Table B has 5 records. Now while inner joining Table A and Table B, there is one duplicate on the joining column in Table B (i.e. Table A has 5 unique records, but Table B has 4 unique values and one redundant value). What will be record count of the output?

- 5 records will be returned. 4 records will be corresponding to the unique entries in both the table. Added to it, even the duplicate row (in B) will find the corresponding identifier in A thus the 5 th record will be there.eg.

First table

id	entry
1	10
2	20
3	30
4	40
5	50

Second table

id	entries
1	100
2	200
3	300
4	400
4	400

Resulting table:

id	entry	entries
1	10	100
2	20	200
3	30	300
4	40	400
4	40	400

#18. What is the difference between WHERE clause and HAVING clause?

- Where clause is used to put certain conditions on columns to fetch desired rows. Having is also used to put conditions on selection of rows, but it uses aggregator functions to put conditions. Eg, select * from table_name where column=condition; select * from table_name group by column2 having sum(column)=condition;
- Having clause is mostly used with group by function.